新时代人民日报分词语料库构建、性能及应用(二)
——深度学习自动分词模型构建

黄水清[1,2] 王东波[1,2]
[1] 南京农业大学信息科学技术学院 南京 210095 [2] 南京农业大学领域知识关联研究中心 南京 210095

摘要:[目的/意义]在新时代人民日报分词语料库的基础上构建的深度学习自动分词模型,不仅有助于为高性能分词模型的构建提供经验,也可以借助具体的自然语言处理研究任务验证深度学习相应模型的性能。[方法/过程]在介绍双向长短时记忆模型(Bi-LSTM)和双向长短时记忆与条件随机场融合模型(Bi-LSTM-CRF)的基础上,阐明汉语分词语料预处理、评价指标和参数与硬件平台的过程、种类和情况,分别构建 Bi-LSTM 和 Bi-LSTM-CRF 汉语自动分词模型,并对模型的整体性能进行分析。[结果/结论]从精准率、召回率和调和平均值 3 个指标上看,所构建的 Bi-LSTM 和 Bi-LSTM-CRF 汉语自动分词模型的整体性能相对较为合理。在具体性能上,Bi-LSTM 分词模型优于 Bi-LSTM-CRF 分词模型,但这一差距非常细微。

关键词:新时代人民日报分词语料 语料库 自动分词 深度学习 Bi-LSTM Bi-LSTM-CRF
分类号:G255.1
DOI:10.13266/j.issn.0252-3116.2019.23.001

1 引言

对于现代汉语文本来说,分词是后续文本处理及探究的基础。没有精准而高效的自动分词模型,现代汉语的词性标注、实体识别、句法分析和机器翻译等都不能有效地进行。自动分词模型的构建一方面取决于高质量人工构建的能够体现时效性的语料库,另一方面也受制于机器学习模型的性能。本文基于自行构建的新时代人民日报分词语料库(New Era People's Daily Segmented Corpus,简称 NEPD)[1],结合相应深度学习模型,探讨基于 NEPD 的现代汉语文本自动分词模型构建问题。这一探究不仅有益于验证深度学习在现代汉语分词上的性能而且有利于后续更加细致和深入地基于深度学习构建高性能的分词模型。NEPD 涵盖了《人民日报》2015 上半年(1-6 月)及 2016 年 1 月、2017 年 1 月、2018 年 1 月共 9 个月的语料,同时进行了人工分词标注,是经过人工加工的精语料[1]。

自深度学习被应用于自然语言处理研究以来,国内外的研究者先后把深度学习应用在了汉语的自动分词探究上,比较有代表性的研究如下:①通过基本的深度神经网络模型探究汉语自动分词的问题。X. Zheng 等[2]利用大规模未标注的数据来改善汉语的内部表示,利用这些改进的表示来加强监督字段的分割,并利用一种训练神经网络的感知式算法,以最小的计算能力改进了中文分词模型的性能。在 4 个自然语言处理任务中,X. Li 等[3]将依赖词的模型和基于神经字符的模型进行比较,发现后者始终优于前者。造成这一结果的原因是基于词的模型更易受到数据稀疏性和词汇不足的影响而产生训练过度拟合问题。该研究的结论启发后续的研究者应该在训练模型的过程中避免过拟合问题的出现。②基于长短时记忆神经网络的自动分词模型构建。张洪刚和李焕[4]提出了 Bi-LSTM 神经网络中文自动分词模型,具体来说,是将字向量应用于 Bi-LSTM 模型实现分词,并在简体和繁体中文数据集上进行了实验。这一探究为进行基于 Bi-LSTM 的汉语自动分词提供了借鉴。在流行的数据集上,J. Ma 等[5]通过实验验证了 Bi-LSTM 模型比复杂的神经网络模型能达到更高的中文分词准确性。但在未登录词这一中文分词的难点上,Bi-LSTM 深度学习模型仍有待改进之处,为解决这一问题,一方面应该对模型进行严格的调优,另一方面应进一步扩大语料库以提高模型的训练性能。通过在微软研究院提供的语料和北京大学人民

作者简介:黄水清(ORCID:0000-0002-1646-9300),教授,博士生导师,E-mail:sqhuang@njau.edu.cn;王东波(ORCID:0000-0002-9894-9550),教授,博士生导师。
收稿日期:2019-11-15 修回日期:2019-12-02 本文起止页码:5-12 本文责任编辑:易飞

日报语料上进行测试，解宇涵[6]提出了一种基于字嵌入的 Bi-LSTM 中文分词模型，并验证了该模型比传统的 HMM 模型在自动分词上更加突出。在继承 LSTM 模型可自动学习特征的基础上，李雪莲等[7]提出了基于门循环单元神经网络的中文分词法，该模型能有效发挥长距离依赖信息的优点，从实验结果上看该方法显著提升了自动分词的性能。通过对不同的语料数据加上人工设定的标识符，姜猛等[8]在所提出的异构处理数据方法基础上，利用 LSTM 模型来对处理过的数据进行训练，实验表明该策略能有效提高分词模型的整体性能。这一研究充分说明了对训练语料进行前期处理的合理性和有效性。在充分挖掘分词对象字位标记特征的基础上，王玮[9]提出了把双向长短期记忆神经网络模型与相应字位标记相融合的自动分词策略，通过与 CRF 等方法对比，实验结果表明 Bi-LSTM 与六字位相融合效果最优。这一研究对于后续探究的启示在于，在条件允许的情况下，对双向 LSTM 进行叠加并融入相应的字位标记能提升分词的整体性能。③基于长短时记忆神经网络与条件随机场相融合的自动分词模型构建。在通过 word2vec 对语料数据的嵌入处理基础上，X. Wang 等[10]把所获取的嵌入特征反馈给 Bi-LSTM，并在输出层添加 CRF，从而构建自动分词模型。与这一方法一样，王梦鸽[11]和薛源[12]也提出了把 Bi-LSTM 与 CRF 相结合的自动分词策略。上述研究所构建的模型的分词结果达到了较好的准确性，本文直接借鉴了这一研究理念和方法。与上一研究相似，在利用分词对象上下文信息的基础上，张子睿和刘云清[13]提出了一种基于长短期记忆神经网络改进的双向长短期记忆条件随机场模型，并通过具体的实验验证了 Bi-LSTM-CRF 模型的整体性能。在中文分词模型 Bi-LSTM-CRF 模型和 seq2seq 模型基础上，刘玉德[14]通过融入注意力机制来对上述模型进行改进，而实验结果表明改进后的分词模型具有更好的分词性能。根据对上述 3 个层次上自动分词相应研究的综述，可以归纳如下两个方面的特征：一方面，上述探究不仅有效使用了深度学习的相应模型，还把深度学习模型与其他机器学习模型进行了融合；另一方面，在发挥深度学习模型优势的同时，也把相应的分词对象的特征添加到了自动分词模型构建当中。

在上述国内外探究的基础上，基于新时代人民日报分词语料，结合 Bi-LSTM 和 Bi-LSTM-CRF 模型，采用十折交叉验证法，本文构建了相对应的深度学习自动分词模型，并对模型的整体性能进行了评价。整个基于深度学习的自动分词模型构建思路如下：首先，通过文献调研分析目前主流的深度学习应用于汉语自动分词的状况，并确定具体所使用的深度学习模型，同时对所选取的模型进行相应的特征和性能分析。其次，结合深度学习模型的特性，对所选取的新时代人民日报语料按照深度学习训练和测试的要求进行预处理，并进行字嵌入的生成。最后，基于所选取的 Bi-LSTM 和 Bi-LSTM-CRF 深度学习模型，构建面向新时代人民日报语料的自动分词模型，并在所选定的模型参数上对所构建的模型进行细致而全面的性能计算和评估。

2 自动分词深度学习模型介绍

从汉语自动分词的任务上看，自动分词是一个典型的线性序列任务。结合已有的相关研究，根据深度学习相应模型的特征，在深度学习自动分词模型的构建中，本文主要基于 Bi-LSTM 和 Bi-LSTM-CRF 这两个具体的深度学习模型完成的。Bi-LSTM 和 Bi-LSTM-CRF 模型的具体特征如下所述：

2.1 Bi-LSTM 模型

在整个深度学习系列模型中，循环神经网络（recurrent neural network，RNN）[15]是一类用于序列标记的人工神经网络，因此该类深度学习模型特别适用于自动分词、词性标注和实体识别等自然语言处理相应的探究任务上。从该模型的理论上来说，RNN 能够学习不同自动分词字特征之间长期的依赖关系属性，但在自动分词模型训练的过程中随着时间序列的推移，RNN 自动分词的深度不断加深，当 RNN 自动分词的层数达到一定的临界值的时候，容易使梯度下降坡度呈指数减少或指数增大，从而导致梯度消失和梯度爆炸现象的出现。而 LSTM 的出现在一定程度上有效地解决了 RNN 的这一历史遗留问题。对于汉语自动分词来说，LSTM 通过实现与记忆单元（memory cell）的结合，并引入门（gate）控制器来控制自动分词模型训练过程中历史信息的保留和丢弃。常规来说，一个 LSTM 神经网络神经元包含一个记忆单元和 3 种门，对于自动分词来说，分别是分词相应信息的输入门（input gate）、分词相应信息的输出门（output gate）和分词相应信息的遗忘门（forget gate）。这 3 种门分别用于控制分词相应信息的输入信息、输出信息和记忆单元中信息的保留或丢弃，从而可以更有效地记忆构建自动分词模型所需的相应分词信息。LSTM 记忆单元的计算公式如下[16]：

$$i_t = \sigma(W_i h_{(t-1)} + U_i x_t + b_i) \quad \text{公式(1)}$$

$f_t = \sigma(W_f h_{(t-1)} + U_f x_t + b_f)$　　公式(2)

$o_t = \sigma(W_o h_{(t-1)} + U_0 x_t + b_o)$　　公式(3)

$c_t = f_t \odot c_{(t-1)} + i_t \odot \tanh(W_c h_{(t-1)} + U_c x_t + b_c)$　　公式(4)

$h_t = o_t \odot \tanh(c_t)$　　公式(5)

对于汉语自动分词模型构建来说，前3个公式中的 i_t、f_t 和 o_t 分别表示的是t时刻的自动分词输入控制门，自动分词遗忘控制门和自动分词输出控制门。最后一个公式中的 c_t 表示的是t时刻与自动分词相关的记忆单元向量。U_i, U_f, U_c, U_o 分别是汉语自动分词的字输入序列 $\{x_0, x_1, \cdots, x_t, x_{t+1}, \cdots\}$ 和各个汉语分词控制门之间的连接权重矩阵，并且是汉语自动分词控制门和隐藏状态h之间的连接权重矩阵。b_i, b_f, b_c, b_o 分别对应了自动分词训练模型中的偏置向量。对于汉语自动分词模型构建来说，通过对记忆单元和控制门的引入，LSTM在一定程度上解决了RNN难以有效获取长度跨度比较大的汉语词汇间的字与字特征的问题。Bi-LSTM模型是拥有两个相反方向LSTM并行层的双向LSTM神经网络，能够同时存储来自两个方向的与汉语自动分词相关的信息。图1是基于NEPD语料的Bi-LSTM汉语自动分词模型架构图：

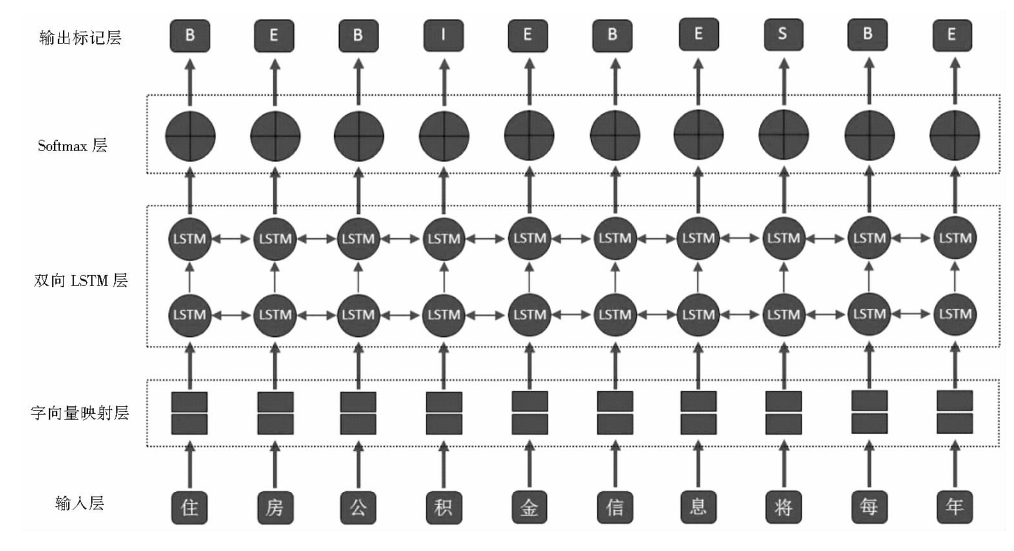


图1　基于Bi-LSTM的汉语自动分词模型架构

从图1可以看出，基于Bi-LSTM的汉语自动分词模型框架共包含5层，从下往上依次为第1层至第5层。第1层是新时代人民日报分词语料的输入层，即将语料以字为单位逐一输入。第2层为新时代人民日报分词语料库中的汉字的字向量映射层，采用分布式表示的方式将每一个语料库中的汉字均转化为128维的字向量，以便于神经元进行汉字特征的特征提取与计算。第3层为双向LSTM神经网络层，从图1可以看出，在汉语自动分词模型构建中，双向LSTM神经网络层拥有两个相反方向并行层的LSTM，可以同时完成从前向后和从后向前对分词语料中的字与字之间的特征进行有效提取与充分训练。第4层为自动分词模型构建的Softmax激活函数层。由于汉语自动分词模型构建涉及到B、I、E、S 4种标签的标注，因此使用维度为4的softmax激活层来进行4种标签的概率预测，以求出可能性最大的汉语自动分词结果输出标签。最后一层是汉语自动分词的标记输出层，经softmax计算后，所得到的每个汉字概率值最大的分词标签将在这一层输出。

2.2　Bi-LSTM-CRF 模型

在汉语自动分词模型构建中，尽管通过Bi-LSTM模型可以获得较好的分词序列效果，但是对于中文自动分词这类输出标签之间存在较强依赖关系的序列标记问题，由于softmax激活函数只能考虑当前汉语字分布状态的特征，不能有效关联汉语字的前后特征，并实现针对汉语自动分词的联合标签概率的预测，因此Bi-LSTM模型下的汉语自动分词模型性能将会受到影响。为了解决汉语自动分词的这一问题，基于Bi-LSTM-CRF[16]构建汉语自动分词模型就应运而生。Bi-LSTM-CRF模型主要是通过去掉Bi-LSTM模型中的Softmax层，代之以CRF线性层而实现把Bi-LSTM模型与CRF模型融合在一起。这一模型组合在完成构建汉语自动分词模型的过程中，不仅保留了Bi-LSTM能够同时考虑当前汉语字的上下文信息的特性，而且还通过CRF层计算整个汉语自动分词观察序列状态标记的联合条件概率分布。图2是基于新时代人民日报分词语料的

Bi-LSTM-CRF 自动分词模型结构图。模型框架同样也包括 5 层,不同的是第 4 层由 Softmax 层变为条件随机场(CRF)层,使得模型在概率计算过程中可以考虑原本相互独立的汉语分词输出标签之间的前后依赖关系,以便于输出最优的分词标签序列。

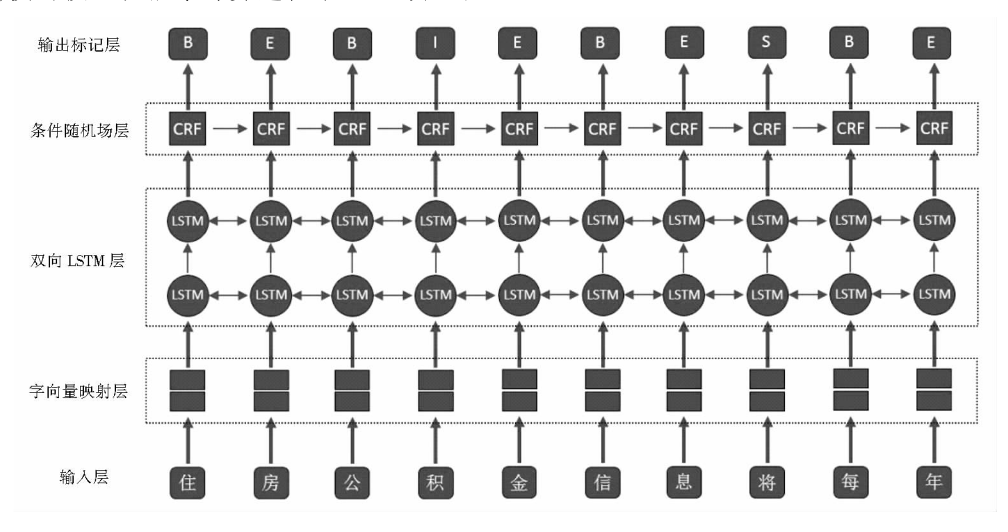


图 2 基于 Bi-LSTM-CRF 的汉语自动分词模型架构

基于上述两个模型所构建的新时代人民日报语料自动分词模型的特点如下:一方面,完全通过基于深度学习模型获取字与字之间构成词汇的特征实现汉语自动分词模型的构建,不涉及到任何人为特征的添加;另一方面,在新时代人民日报语料的基础上,充分利用了 Bi-LSTM 能够解决模型在训练过程中出现的梯度消失和梯度爆炸问题这一特性,从而从整体上确保了所构建汉语自动分词模型的性能。

3 深度学习分词模型性能分析

在这一部分主要对新时代人民日报语料的预处理、评价指标和模型参数与硬件平台进行了介绍,并在介绍的基础上重点对所构建的 Bi-LSTM 和 Bi-LSTM-CRF 汉语自动分词模型的整体性能进行了细致和全面的分析。所构建的新时代人民日报语料深度学习自动分词模型的创新性主要如下:一方面,实现了深度学习自动分词模型与体现时代特征的语料的有机结合,从而确保了所构建的自动分词模型能够高效和精准地完成对相关领域文本的自动分词;另一方面,针对多字词,从词的构成成分角度对自动分词的整体性能进行了探究,这为深入分析深度学习模型在自动分词这一研究任务上的性能提供了新的视角。

3.1 语料预处理及评价指标

基于 Bi-LSTM 和 Bi-LSTM-CRF 两种深度学习模型,在新时代人民日报语料上,训练和测试自动分词模型,并对两种模型的性能进行分析。深度学习分词模型构建的流程具体如下:

首先,基于人民日报汉语词汇的字长,在构建深度学习自动分词模型的过程中所使用的标记集由“B,I,E,S”4 个标记构成,“B”“I”和“E”分别代表多字词的首字词、中间字和尾字,而“S”则表示单字词。如果一个词的长度超过了 3 个字,则让“I”循环表示中间出现的字。

其次,从新时代当中选取 2018 年 1 月人民日报语料作为构建分词模型的训练和测试语料。在把人民日报语料转化为深度学习模型可以训练和测试的语料过程中,结合所制定的标记集及标注准则,对人民日报语料实现了训练和测试格式的转换,具体样例如表 1 所示:

表 1 深度学习训练和测试语料样例

编号	训练和测试字序列	标记
1	参	B
2	加	E
3	的	S
4	多	S
5	是	S
6	热	B
7	心	E
8	社	B
9	区	E
10	事	B
11	务	E
12	的	S

在样例表 1 中,由于字序列这一行主要由单字词

和二字词构成，在标记这一列当中，主要使用了 B，E 和 S 这 3 个标记。

最后，对所构建的深度学习分词模型评价仍使用精准率、召回率和调和平均值这 3 个指标。在具体评价的过程中，为了更加细致和全面地评价所构建深度学习分词模型的整体性能，不仅对整体的分词标记进行评价，而且对单一的分词标记逐一地进行评价。

3.2 深度学习模型参数及硬件平台

本文分别使用 Bi-LSTM 模型、Bi-LSTM-CRF 模型进行自动分词模型构建。对于每种模型，均采用十折交叉训练的方式，以排除随机误差对实验结果的影响。Bi-LSTM 模型和 Bi-LSTM-CRF 模型这两个深度学习模型主要由 Embedding 层、双向 LSTM 层和 CRF 层构成。在模型具体训练过程中，LSTM 层数设置为 2，而每个 LSTM 层的隐藏单元数(hidden unit)则设定为 256。为了防止在自动分词模型构建过程中梯度爆炸与消失问题的出现，本文采用梯度裁剪(gradient clipping)技术，并把其值设置为 5.0。在训练的数据输入过程中，每批数据量(batch size)大小设定为 32，而隐藏单元随机删除概率(dropout rate)则设置为 1，相应的学习率(learning rate)设置为 0.001。整个训练模型的字嵌入(word embedding)则通过 gensim 包的 word2vec 进行预训练，向量维度设置为 128 维，而训练周期(Epoch)设置为 200，梯度优化器(Optimizer)为 Adam。为了防止过拟合现象并加快训练速度，在模型训练过程中采用 early stop 模式，当交叉验证集 F 值 10 次不提高时，则停止整个分词模型的训练。

由于神经网络在训练过程中需要涉及大量的并行运算和矩阵计算，在中央处理器(CPU)上开展深度学习任务时，无法提供足够的吞吐量和响应速度。因此，本文使用高性能的 NVIDIA Tesla P40 图形处理器(GPU)进行神经网络的训练，它可提供比 CPU 快 60 倍以上的处理能力，可达到 47 TOPS(万亿次运算/秒)的推理性能。本文使用的计算机配置情况介绍如下：CPU:48 颗 Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz；内存:256GB；GPU:6 块 NVIDIA Tesla P40；显存:24GB；操作系统：CentOS 3.10.0。

3.3 基于 Bi-LSTM 模型的自动分词性能分析

基于训练和测试语料，通过 Bi-LSTM 模型，构建了 10 个自动分词模型，并基于精准率、召回率和调和平均值对 10 个模型的性能进行了评测，具体性能如表 2 所示：

表 2 基于 Bi-LSTM 模型的自动分词性能

模型	评测对象	精准率/%	召回率/%	调和平均值/%
模型 1	B	97.84	98.94	98.39
	E	97.73	98.79	98.26
	I	93.91	85.41	89.46
	S	97.68	97.28	97.48
	所有标记	97.51	97.51	97.51
模型 2	B	97.86	98.89	98.37
	E	97.72	98.75	98.23
	I	94.19	85.76	89.78
	S	97.50	97.24	97.37
	所有标记	97.48	97.48	97.48
模型 3	B	97.81	98.86	98.33
	E	97.69	98.73	98.21
	I	93.95	86.07	89.84
	S	97.60	97.15	97.38
	所有标记	97.47	97.47	97.47
模型 4	B	97.76	98.91	98.34
	E	97.62	98.75	98.18
	I	94.25	84.98	89.38
	S	97.52	97.21	97.36
	所有标记	97.43	97.43	97.43
模型 5	B	97.72	98.84	98.28
	E	97.61	98.72	98.16
	I	94.17	85.28	89.50
	S	97.50	97.18	97.34
	所有标记	97.40	97.40	97.40
模型 6	B	97.82	98.89	98.35
	E	97.67	98.76	98.21
	I	94.19	85.20	89.47
	S	97.50	97.23	97.37
	所有标记	97.45	97.45	97.45
模型 7	B	97.80	98.96	98.38
	E	97.67	98.80	98.23
	I	94.61	85.38	89.76
	S	97.61	97.24	97.43
	所有标记	97.50	97.50	97.50
模型 8	B	97.72	98.95	98.33
	E	97.56	98.80	98.17
	I	94.49	84.52	89.23
	S	97.53	97.17	97.35
	所有标记	97.41	97.41	97.41
模型 9	B	97.88	98.92	98.40
	E	97.75	98.75	98.25
	I	94.06	85.85	89.77
	S	97.58	97.25	97.42
	所有标记	97.52	97.52	97.52
模型 10	B	97.74	98.91	98.32
	E	97.58	98.76	98.17
	I	94.43	84.73	89.32
	S	97.51	97.25	97.38
	所有标记	97.42	97.42	97.42

根据表 2 中所有标记的精准率和召回率，Bi-LSTM 自动分词模型的平均调和平均值为 97.46%，其中最高的调和平均值为 97.52%。通过对 B、E、I 和 S 4 个标记性能的分析，影响所有标记平均调和平均值性能提升的原因在于中间字的整体性能较差。在 Bi-LSTM 自动分词模型中，单字词的平均调和平均值为 97.39%，相较于所有标记的平均调和平均值来说降低了 0.07%。单字词的最高调和平均值为 97.48%，超过了所有标记的平均调和平均值，而比所有标记的最高调和平均值低 0.07%。为了更加细致而全面地对 Bi-LSTM 自动分词模型的性能进行分析，对多字词的首字、中间字和尾字进行了评估。多字词的首字平均调和平均值为 98.35%，比所有标记的平均调和平均值高 0.89%，而首字的最高调和平均值则达到了 98.40%。由于多字词的总长度一定程度上是由中间字的个数决定的，对于长度比较大的词汇来说，中间字的标注性能整体上就会较差。Bi-LSTM 自动分词模型的中间字的平均调和平均值为 89.55%，最高的调和平均值为 89.84%，而最低的则仅为 89.23%。多字词的尾字的平均调和平均值为 98.21%，比所有标记的平均调和平均值高 0.25%，其中最高的调和平均值为 98.26%。

3.4 基于 Bi-LSTM-CRF 模型的自动分词性能分析

为了验证 Bi-LSTM 模型在融入 CRF 解决输出结果的偏置上的整体性能，在人民日报分词的训练和测试语料上，完成了对 Bi-LSTM-CRF 分词模型的构建并对所构建模型进行了评测。Bi-LSTM-CRF 分词模型的精准率、召回率和调和平均值具体见表 3。

从表 3 中可以计算出，所有标记的 Bi-LSTM-CRF 分词模型的平均调和平均值为 97.43%，最高调和平均值为 97.49%，比所有标记的 Bi-LSTM 分词模型的平均调和平均值和最高调和平均值均低了 0.03%。Bi-LSTM 分词模型的单字词的平均调和平均值为 97.37%，最高调和平均值为 97.47%，仅比 Bi-LSTM 分词模型的单字词平均调和平均值和最高调和平均值低了 0.02% 和 0.01%。多字词的首字的 Bi-LSTM-CRF 分词模型的平均调和平均值为 98.34%，最高调和平均值为 98.38%，与 Bi-LSTM 分词模型的多字词的首字平均调和平均值和最高调和平均值相比低了 0.01% 和 0.02%。而对于多字词的中间字的标注，相较于 Bi-LSTM 模型，融入了 CRF 的 Bi-LSTM 模型整体性能有略微的下降，平均调和平均值为 89.28%，而最高调和平均值为 89.62%，比 Bi-LSTM 模型的平均调和平均值

表 3 基于 Bi-LSTM-CRF 模型的自动分词性能

模型	评测对象	精准率/%	召回率/%	调和平均值/%
模型 1	B	97.81	98.95	98.38
	E	97.68	98.82	98.25
	M	94.79	84.22	89.19
	S	97.48	97.46	97.47
	所有标记	97.49	97.49	97.49
模型 2	B	97.84	98.91	98.37
	E	97.68	98.76	98.22
	M	94.14	85.39	89.55
	S	97.51	97.22	97.36
	所有标记	97.46	97.46	97.46
模型 3	B	97.77	98.88	98.32
	E	97.64	98.75	98.20
	M	94.48	84.93	89.45
	S	97.44	97.28	97.36
	所有标记	97.43	97.43	97.43
模型 4	B	97.77	98.83	98.30
	E	97.62	98.68	98.15
	M	93.42	84.94	88.98
	S	97.46	97.18	97.32
	所有标记	97.36	97.36	97.36
模型 5	B	97.73	98.79	98.26
	E	97.62	98.67	98.14
	M	94.03	84.88	89.22
	S	97.37	97.26	97.32
	所有标记	97.36	97.36	97.36
模型 6	B	97.79	98.88	98.33
	E	97.65	98.73	98.19
	M	93.80	84.84	89.10
	S	97.46	97.21	97.33
	所有标记	97.41	97.41	97.41
模型 7	B	97.84	98.90	98.37
	E	97.68	98.74	98.21
	M	93.98	85.64	89.62
	S	97.58	97.20	97.39
	所有标记	97.47	97.47	97.47
模型 8	B	97.77	98.87	98.32
	E	97.62	98.72	98.17
	M	93.79	84.70	89.01
	S	97.46	97.23	97.35
	所有标记	97.38	97.38	97.38
模型 9	B	97.87	98.90	98.38
	E	97.72	98.75	98.23
	M	94.05	85.37	89.50
	S	97.49	97.25	97.37
	所有标记	97.48	97.48	97.48
模型 10	B	97.75	98.89	98.32
	E	97.61	98.75	98.17
	M	94.31	84.57	89.17
	S	97.47	97.30	97.38
	所有标记	97.41	97.41	97.41

和最高调和平均值分别低了0.27%和0.22%。在多字词的尾字上,Bi-LSTM-CRF分词模型的平均调和平均值和最高调和平均值分别为98.19%和98.25%,分别比Bi-LSTM模型的平均调和平均值和最高调和平均值低了0.02%和0.01%。

在汉语自动分词的任务上,通过对比基于Bi-LSTM和Bi-LSTM-CRF所构建自动分词模型的性能得出如下认识。首先,在汉语自动分词这一任务上,Bi-LSTM和Bi-LSTM-CRF这两个深度学习模型的整体性能差距非常小。从整体性能的评估到具体标记的评测,这一差距基本上维持在0.01%到0.27%之间。其次,对于汉语自动分词模型的构建,Bi-LSTM-CRF在目前的参数设置上,其对于结果输出的位置偏移性这一问题解决优势并没有得到充分的体现。最后,对于汉语词汇长度比较大的分词任务来说,Bi-LSTM和Bi-LSTM-CRF模型性能整体上均不是太突出。对于这一点,多字词的中间字的调和平均值均较低就是有力证明。

4 结语

为了探究基于深度学习构建自动分词模型的整体性能和状况,本文在新时代人民日报分词语料的基础上,结合Bi-LSTM与Bi-LSTM-CRF深度学习模型,对自动分词模型的构建进行了系统的探究。基于Bi-LSTM和Bi-LSTM-CRF这两种深度学习模型,本文完成了深度学习汉语自动分词模型的构建,并从宏观和微观两个维度上对比了两种不同模型的整体分词性能。囿于语料的规模和模型训练的时间,在构建本文的深度学习自动分词模型的过程中对所涉及的参数只进行了简单的测验,因此在后续的探究中不仅要强化对模型参数在各个维度上的验证,而且要构建新的深度学习自动分词模型,以便于从多角度、多维度上探究深度学习在汉语自动分词这一研究任务上的性能。

参考文献:

[1] 黄水清,王东波.新时代人民日报分词语料库构建、性能及应用(一)——语料库构建及测评[J].图书情报工作,2019,63(22):5-12.

[2] ZHENG X,CHEN H,XU T. Deep learning for Chinese word segmentation and POS tagging [C]// YAROWSKY D, BALDWIN T, KORHONEN A, et al. Proceedings of the 2013 Conference on empirical methods in natural language processing. Washington:Association for computational Linguistics, 2013: 647-657.

[3] LI X, MENG Y, SUN X, et al. Is word segmentation necessary for deep learning of Chinese representations? [J]. [2019-11-10]. https://arxiv.org/abs/1508.01991v1.

[4] 张洪刚,李焕.基于双向长短时记忆模型的中文分词方法[J].华南理工大学学报(自然科学版),2017(3):61-67.

[5] MA J, GANCHEV K, WEISS D. State-of-the-art Chinese word segmentation with Bi-LSTMs[C]// RILOFF E, CHIANG D, HOCKENMAIER J, et al. Proceedings of the 2018 conference on empirical methods in natural language processing. Belgium:Association for Computational Linguistics, 2018:4902-4908.

[6] 解宇涵.基于深度学习的中文分词模型应用研究[D].重庆:重庆大学,2017.

[7] 李雪莲,段鸿,许牧.基于门循环单元神经网络的中文分词法[J].厦门大学学报(自然科学版),2017,56(2):237-243.

[8] 姜猛,王子牛,高建瓴.基于异构数据联合训练的中文分词法[J].电子科技,2019,32(4):33-36.

[9] 王玮.基于Bi-LSTM-6Tags的智能中文分词方法[J].计算机应用,2018,38(S2):112-115.

[10] WANG X, WANG M, ZHANG Q. Realization of Chinese word segmentation based on deep learning method[C]// Green Energy and Sustainable Development I. Proceedings of the international conference on green energy and sustainable development. Chongqing: AIP Publishing, 2017:1-6.

[11] 王梦鸽.基于深度学习中文分词的研究[D].西安:西安邮电大学,2018.

[12] 薛源.基于深度学习算法的中文分词的研究[J].计算机产品与流通,2019(5):202.

[13] 张子睿,刘云清.基于BI-LSTM-CRF模型的中文分词法[J].长春理工大学学报(自然科学版),2017,40(4):87-92.

[14] 刘玉德.基于深度学习的中文分词方法研究[D].广州:华南理工大学,2018.

[15] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088):533-536.

[16] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. [2019-11-10]. http://arxiv.org/abs/1508.01991v1.

作者贡献说明:

黄水清:提出相关概念及整体研究思路,修订完稿;

王东波:数据处理及初稿撰写。

Construction, Performance and Application of New Era People's Daily Segmented Corpus (Ⅱ)
——Constructing Automatic Word Segmentation Model of Deep Learning

Huang Shuiqing[1,2] Wang Dongbo[1,2]

[1] College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095

[2] Research Center for Correlation of Domain Knowledge, Nanjing Agricultural University, Nanjing 210095

Abstract: [**Purpose/significance**] On the basis of the new era People's Daily(NEPD) word segmentation corpus, the construction of the automatic word segmentation model of deep learning not only can help to provide relevant experience for the construction of high-performance word segmentation model, but also can verify the performance of the corresponding model of deep learning through specific natural language processing tasks. [**Method/process**] Based on the introduction of Bi-directional Long Short-Term Memory (Bi-LSTM) and Bi-directional Long Short-Term Memory with conditional random field (Bi-LSTM-CRF), this paper expounded the process, type and situation of Chinese word segmentation preprocessing, the evaluation indexes and parameters and hardware platform, the Bi-LSTM and Bi-LSTM-CRF Chinese automatic word segmentation models were constructed respectively, and the overall performance of the models was analyzed. [**Result/conclusion**] The overall performance of the Bi-LSTM and Bi-LSTM-CRF Chinese automatic word segmentation model is relatively reasonable from the three indexes of precision, recall and F value. In terms of specific performance, Bi-LSTM word segmentation model is superior to Bi-LSTM-CRF word segmentation model, but the difference is very small.

Keywords: new era People's Daily segmented corpus segmented corpus automatic word segmentation deep learning Bi-LSTM Bi-LSTM-CRF